Behavioral Housing Search Choice Set Formation: A Spatial Hazard-based Screening Model

Taha H. Rashidi¹ and Abolfazl (Kouros) Mohammadian²

Abstract
The housing search process, a topic of interest to both practitioners and researchers, starts with an alternative formation and screening practice. Due to the limitation of cognitive capacity, household members at this level evaluate potential alternatives based on many factors, such as lifestyle, preferences, and so on, to form a manageable choice set. This article attempts to provide a detailed study of this screening and filtering practice to develop a modeling framework that can replicate the choice set formation process. In order to show the potential of the method, one prospective decision criteria—the average desired commute to work distance—is considered the potential attribute that the household evaluates for feasible housing alternatives. It is postulated that alternatives will only be included in the choice set if the average work distance satisfies the household distance threshold. This article explores the viability of using proportional hazard models in the housing search process. Some of the specifications of hazard-based models that are typically used on temporal data are examined on average work distance. Several household socio-demographic attributes from eight waves of the Seattle Metropolitan Area’s Puget Sound Transportation Panel (PSTP) are utilized for model estimation, along with built environment variables, characteristics of the supply side of the market, and

¹ Department of Civil Engineering, University of Toronto, Toronto, Canada
² Department of Civil and Materials Engineering, University of Illinois, Chicago, IL, USA

Corresponding Author:
Taha H. Rashidi, Department of Civil Engineering, University of Toronto, Toronto, Canada M5S1A4
Email: taha.hosseinrashidi@utoronto.ca
several other economic indicators. The approach presented in this article provides a remedy for the large choice set problem typically faced in discrete choice modeling.

**Keywords**
land use and urban form, urban and regional spatial structure, spatial structure, residential location, neighborhood, human spatial structure, housing, urban and regional issues, policy and applications

**Introduction**
The spatial location decision-making process has become a topic of interest in many fields, including transportation, urban planning, psychology, and other related disciplines. Since the early introduction of the discrete choice paradigm (McFadden 1974), the individual’s alternative selection behavior has been primarily modeled using the discrete choice modeling approach. However, the prediction potential of a discrete choice model and the accuracy of its parameter estimates are highly dependent on the choice set composition.

This study proposes a behavioral method for housing search choice set formation. Previously, Thill and Horowitz (1997) discussed an approach to the context of destination choice, in which they used travel time for alternative screening and assumed that it was an unobservable random variable that did not depend on any observable attribute of travelers. In the current article, after a decision maker became active in the housing market, it was assumed that residential location choice process started with an alternative evaluation and screening practice. People scanned their alternatives first, then filtered them based on their priorities, lifestyle, preferences, budget, and perceived utilities. Finally, among the filtered alternatives, the most desired option with the highest utility was chosen.

The understanding of spatial decision processes is improving as a result of increasing emphasis and discussions regarding new modeling approaches, techniques, and issues related to spatial decision, and, particularly, housing search behavior (Fischer, Nijkamp, and Papageorgiou 1990; Cascetta, Pagliara, and Axhausen 2007).

While several factors affect the selection of housing alternatives and the spatial choice decision mechanism (e.g., property value, commute distance, school quality, safety, tax rate, etc.), in order to show the practicality of the approach, an influential factor such as commute distance, known as an essential variable in residential selection behavior, is considered in the screening process model of this study (Kim 1992; Van Ommeren, Rietveld, and Nijkamp 1997; Rashidi, Mohamamdian, and Koppelman 2011). Then, a hazard-based modeling approach which is traditionally used for modeling duration is introduced and utilized as the mathematical toolbox for modeling. In that model, commute distance is considered a continuous and nonnegative random variable.
The analogy between duration and commute distance (spatial duration) is justifiable from a mathematical perspective. Furthermore, estimation, specification, tests, and diagnostic issues can be easily handled using the duration modeling methodology. This analogy has been intermittently announced in some research areas with claims that the spatial duration model can be a useful toolbox for analyzing residential location search behavior (Odland and Ellis 1992; Diggle 1983; Boots and Getis 1988). Nonetheless, some conceptual difficulties in the interpretation of spatial duration models have not yet been sufficiently addressed in this emerging field. At the same time, the advantages of applying the longitudinal framework in the context of spatial duration have not been appropriately discussed.

In short, the two major contributions of this article are the introduction of an innovative choice set formation in an application of the housing search problem and the exploration of the viability of using proportional hazard formulation in this application. The remainder of the article is structured as follows. First, a brief literature review is presented and the study approach is discussed. Model derivation and the mathematical formulations of the system of equations are presented next. The data sets used in this study are then explained, and their key variables are discussed. Following that, experimental results of different steps of the parameter estimation process are presented. Conclusions and future research directions are discussed in the final section.

**Background**

**Choice Set Formation**

Despite the great effort to improve the theory and applications of the discrete choice models (Thill 1992), the importance of the choice set composition on parameter estimation has seldom been examined (Ben-Akiva and Lerman 1985; Timmermans and Golledge 1990). In the literature, there have been two extreme approaches to selecting the set of alternatives: first, randomly selecting a finite number of alternatives from the universal choice set as defined by Ben-Akiva and Lerman (1985), and second, considering all plausible alternatives (Salomon and Ben-Akiva 1983; Thill and Horowitz 1991). Both approaches can raise serious concerns. Although the inclusion of all possible alternatives may appear to be a conservative approach, it can be unrealistic as it assumes that decision makers have perfect knowledge of all possible alternatives. This approach can result in assigning nonnegative selection probability to some alternatives that may not be otherwise known or available to the decision maker (Richards and Ben-Akiva 1975; Domencich and McFadden 1975). On the other hand, a random selection of fewer alternatives for the choice set by stratified sampling or other similar approaches can result in bias and possibly inaccurate parameter estimation.

Regardless of the type of the method used for choice set composition, it is critical to employ an appropriate filtering/screening method. An accurate estimation of discrete choice model parameters and a correct prediction of choices (Manski
1977) are conditional on correct information about the choice sets. It is worth noting that, as long as the individual’s true alternatives are considered in the choice set, under specific conditions and constrains (Swait and Ben-Akiva 1987), a logit model can accurately estimate the parameters (McFadden 1978). Similarly, Manski mentioned in his article that, under specific conditions, the approximation of the true choice set by a systematically selected subset does not compromise the consistency of the choice model estimates (Manski 1977). However, this would not be true in the case of a misspecified choice set, which could result in erroneous parameter estimates (Lerman 1985). Although such effects are marginal, behavioral insight into the methods of choice set formation may considerably improve the accuracy of the model and parameter estimates, whereas the absence of this insight could be problematic (Burnett and Hanson 1979, 1982).

Srinivasan (1987) introduced three levels in screening the alternatives and finding the final choice set: awareness set, evoked set, and choice set. The term evoked set was introduced in an earlier study by Howard, who originally introduced the term in 1963. According to his model, the awareness set consists of all alternatives of which the consumer is aware. This set is then filtered to the evoked set, which is a subset of the awareness set and consists of those alternatives that meet certain criteria for further consideration. Finally, the choice set is a subset of the evoked set, in which the few alternatives include the final choice, which is the immediate group of alternatives before making a decision. Shocker et al. (1991) employed the term consideration set for evoked set, which was originally introduced in a study by Wright and Barbour in 1977. Other than concepts of the universal set, awareness set, consideration and evoked set that were the first to appear in the literature, several other types of sets have been introduced in marketing and tourism research, including opportunity set (Woodside and Sherrell 1977), action set (Spiggle and Sewall 1987), insert set (Narayana and Markin 1975), foggy set (Brisoux and Laroche 1981), inaction set (Spiggle and Sewall 1987), hold set (Brisoux and Laroche 1981), reject set (Brisoux and Laroche 1981), inept set (Narayana and Markin 1975), and unavailable aware set (Woodside and Lysonski 1989).

A number of papers in the tourism literature have looked at conceptualization choice set formation. These studies focus on the evolution of vacation destinations and/or plans in the process of choice set formation (Hong et al. 2006; Perdue and Meng 2006; Decrop 2010). Nonetheless, such studies are still in their early stages, just as the current article is exploring choice set modeling possibilities for housing search behavior. Other than the different definitions for the choice set, various solutions have been introduced to deal with the choice set problem. Willumsen and Ortuzar (2001) listed three ways to tackle the choice set problem available in the literature:

1. Rule-based heuristic or deterministic choice set generation methods
2. Simply surveying the individuals regarding their opinions on the feasible alternatives
3. Application of random choice sets.
Lerman (1984) proposed a two-step process; initially, the probability distribution function across all possible choice sets is defined, and then, conditional on the specific choice set, the choice probability of each alternative in the choice set is defined. This study focuses on the first step of the two-step approach by introducing a hazard-based approach that is shown to result in a realistic random choice set. The second step of this approach is kept as a remaining research task (Rashidi, Auld, and Mohammadian 2012).

**Spatial Hazard-based Models**

Cox discussed the basics of duration models as an analysis of exponentially distributed lifetimes were discussed in 1959. These statistical techniques were generalized in another greatly referenced paper of his in 1972. In his later paper, he discussed a satisfying analysis method for failure times, such as the length of time a person is alive. Generally speaking, in a duration model, the time to reach failure, loss, or censoring is observed for each individual in the population. Later, in 1984, Cox and Oakes published a book titled, *Analysis of Survival Data* in which they provided a comprehensive study about hazard models and related topics. This book is one of the most important references in the duration modeling literature. Since these early discussions about hazard-based and duration models, an extensive amount of discussions and applications of these models has been presented (see, e.g., Han and Hausman 1990; Rashidi and Mohammadian 2011). Nonetheless, it has been only recently that discussions about the conceptual equivalence between spatial duration models, in which duration is substituted with distance, can be found in the literature with their temporal counterparts (Waldorf 2003; Carruthers et al. 2009).

In one of the early studies on spatial duration, Rogerson, Weng, and Lin (1993) used the spatial duration approach to model the observed distances between the location of parents and locations of their adult children. In a similar study, Esparza and Krmenec (1994, 1996) modeled the distances between producer service providers and their clients in a spatial duration context. Pellegrini and Grant (1999) expanded the application of spatial hazard models to distances in more abstractly defined spaces—for example, nearest neighbor distances, rather than simple distances in physical space. Such applications of spatial hazard-based models are still in their infancy; only very basic specifications of the longitudinal hazard models are utilized in the spatial duration models. This article attempts to explore some of the more advanced specifications of parametric hazard-based models in a spatial duration application to estimate the household decision on work distance.

Commute distance, known as a key variable in housing search behavior, is utilized in this study to construct the choice set of the housing decision. It is well known that the distance between residence and job location has a significant impact on residential location choice behavior. The correlation between job and residential
locations has been studied extensively in the literature, in most of which commute
distance is distinguished as the critical factor (Van Ommeren, Rietveld, and Nijkamp
1997, 1999; Rashidi, Mohamamdian, and Koppelman 2011). In an interdisciplinary
study, Waddell (1996) modeled the interactions between workplace, residential
mobility, tenure, and location choices in a nested Logit framework. However, there
is a gap in the literature for studying the commute distance as the dependent
variable of a spatial duration model. This article attempts to address this missing
research gap.

**Model Formulation and Methodology**

In this section, an introduction to the parametric hazard-based models is presented.
Cox, who pioneered the area of hazard models in 1959, presented the early versions
of hazard models with the Weibull baseline hazard, which will be further discussed
in this section. Since then, the Weibull function has been frequently used in the
duration modeling context by many other researchers. It should be noted that in all
of the formulations discussed in this section, duration can be replaced with distance
without losing the generality.

The length of a spell for a subject (e.g., a household) is translated in the hazard
formulation as a continuous random variable $T$ with a cumulative distribution
function (CDF), $F(t)$, and probability density function (PDF), $f(t)$ where $t$ is the
elapsed time since entry to the state at time 0. The survival function is defined as
$1 - F(t)$ and is also known as the failure function. In the mathematical context, the
failure function can be written as:

$$
Pr(T \leq t) = F(t).
$$

Therefore, the survival function can be written as:

$$
Pr(T > t) = 1 - F(t) = S(t).
$$

PDF, which is the slope of the failure function, is

$$
f(t) = \lim_{\Delta t \to 0} \frac{Pr(t + \Delta t \geq T \geq t)}{\Delta t} = \frac{\partial F(t)}{\partial t} = -\frac{\partial S(t)}{\partial t},
$$

where $\Delta t$ is a very small *infinitesimal* interval of time.

The hazard rate can be defined then as the probability of leaving in the interval
$(t, t + \Delta t]$, conditional on survival up to time $t$:

$$
\lambda(t) dt = Pr(t + \Delta t \geq T \geq t | T \geq t) = f(t) dt = \frac{S'(t) dt}{S(t)}.
$$

where $\lambda(t)$ is the probability of failure for individual $i$ given that it has survived until
time $T$; $f(t)$ is failure PDF and $S(t)$ is the survival function.
The survival function can be calculated using equation (4) as:

\[
S(t) = \exp \left[ -\int_0^t \lambda(u) \, du \right].
\]

When the impacts of person-specific covariates, like sociodemographic attributes, built-environment variables, and macroeconomic factors are included in a hazard model, it is called a proportional hazard model (Cox 1959), which can be formulated as:

\[
\lambda_i(t) = \lambda_0(t) \exp(-\theta_i X_i),
\]

where \( \lambda_0(t) \) is the baseline hazard function, which only depends on \( t \) (but not \( X \)). It represents the pattern of duration dependence and is assumed to be common among all households. \( \theta_i \) is the coefficient of the covariates. The exponential part of equation (6) scales the baseline hazard function depending on the covariates. The most important property of the proportional hazard is that absolute differences in \( X \) imply proportionate differences in the hazard values at a specific time \( t \). In mathematical terms:

\[
\frac{\lambda_i(t)}{\lambda_j(t)} = \frac{\lambda_0(t) \exp(-\theta_i X_i)}{\lambda_0(t) \exp(-\theta_j X_j)} = \exp[-\theta_i (X_i - X_j)].
\]

Consequently, the proportionate change in the hazard function can be shown by a unit change in covariate \( X \).

Unlike the nonnegative part for the covariates, which is always used in an exponential form, the baseline hazard part can take several shapes among which Weibull and log-logistic are the most well-known functions:

Weibull baseline hazard:

\[
\lambda_i(t) = \gamma t^{\gamma-1} \exp(-\theta_i X_i),
\]

Log-logistic baseline hazard:

\[
\lambda_i(t) = \frac{\beta (t/\alpha)^{\beta-1}}{1 + (t/\alpha)^\beta} \exp(-\theta_i X_i),
\]

where \( \gamma \) is the shape parameter of the Weibull distribution, \( \alpha \) and \( \beta \) are scale and shape parameters of log-logistic distribution, \( X \) denotes explanatory variables, \( \theta_i \) is the vector of parameters. Using the same definitions, the survival function with Weibull or log-logistic assumption for the baseline hazard can be shown as:

Weibull baseline hazard:

\[
S_i(t) = e^{-\gamma t^{\gamma} \exp(-\theta_i X_i)},
\]

Log-logistic baseline hazard:

\[
S_i(t) = \left( 1 + \left( \frac{t}{\alpha} \right)^\beta \right)^{-\exp(-\theta_i X_i)}.
\]
In a mathematical language, the likelihood of failure in accepting a price or a distance while examining different alternatives is equal to the hazard of failure to accept the alternative times the probability of surviving without accepting it. The likelihood function is used to estimate the model parameters.

Data

The Puget Sound Transportation Panel (PSTP) is used as the primary source of data in this study. The PSTP is a panel data for Seattle Metropolitan Area (Murakami and Watterson 1992). Nonetheless, only household observations of the King and Kitsap county areas are used for the modeling practice, due to a need for auxiliary data (e.g., property values, etc.) that were not available for other counties. Of the existing ten waves that the PSTP covered during the last decade of the twentieth century (plus the first two years of the twenty-first century), eight waves are included in this study. The PSTP provides a wide range of variables in the household level, including household sociodemographic attributes. Furthermore, person-level attributes such as home-to-work distances are also provided in the PSTP.

Average household work distance is directly obtained by running queries on the PSTP data. Another important variable, which is included in the models developed in this study, is the property value not included in the PSTP. Land values and house prices are mainly attained from county assessment departments. This information is mainly obtained for property tax preparation purposes and is discarded after a decade or so. Such data were only available from King and Kitsap counties at the TAZ and tract levels. The data retrieved from the two counties (King County Assessment Department 2009; Kitsap County Public Data 2010) were at the very detailed parcel level and subsequently aggregated into the census tract level, then coordinated with the PSTP data using a GIS application.

King and Kitsap counties’ built-environment characteristics are borrowed from an adjunct survey of the PSTP, in which intersection density, different job category counts, transit availability, and many other land-use-related variables in a grid of 150 km² are presented.

Finally, historical macroeconomic data are also merged to the aforementioned data sets. Variables like interest rate, inflation rate, gas price, and unemployment rate are all tested in the models, and their impact on the household decision regarding residential location attributes are examined. Table 1 shows a summary of the independent variables that were utilized in the modeling practice of this study.

The variables in dollar units are adjusted for the inflation value. In other words, in order to keep all prices and income values comparable, the first wave of the PSTP was assumed to be the base year, and dollar values referring to years after the base year were deflated to the base year using the historical inflation rates. The historical thirty-year mortgage rates (HSH Associates Financial Publishers 2009) were also converted to the real mortgage rates by subtracting the nominal values from the corresponding year’s inflation rates.
The eighth variable shown in Table 1, Neighbor_Jobs, represents the number of jobs existing in the neighbor TAZs. This variable can reflect the employment situation in the zones surrounding the past house location. It also shows that the pattern of the residential location moves in the Puget Sound data confirms the sprawling pattern happening in the region. As seen in the Table 1, the average commute distance is increased by 16.4 percent while comparing the last two rows in the table. It is worth noting that the commute distance variable discussed in this article represents the average commute distance of all employed household members. As a result, the residential relocation decision is not limited to the preference of one household member because relocation is a household level decision. Figure 1 also shows the distribution of the observations in the Seattle area.

### Table 1. Descriptive analysis of the dependent and independent variables used in the models

<table>
<thead>
<tr>
<th>ID</th>
<th>Variables’ Name</th>
<th>M</th>
<th>SD</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GasPrice ($)</td>
<td>0.75</td>
<td>0.20</td>
<td>Adjusted gas price in the transaction year</td>
</tr>
<tr>
<td>2</td>
<td>Employed</td>
<td>1.28</td>
<td>0.82</td>
<td>Number of employed people in the household</td>
</tr>
<tr>
<td>3</td>
<td>TotAdult</td>
<td>1.88</td>
<td>0.59</td>
<td>Number of adults in the household</td>
</tr>
<tr>
<td>4</td>
<td>Tot6_17</td>
<td>0.42</td>
<td>0.78</td>
<td>Number of children in the age of 6 to 17</td>
</tr>
<tr>
<td>5</td>
<td>Mortgage</td>
<td>6.98</td>
<td>0.90</td>
<td>Adjusted mortgage rate of the year before the transaction year</td>
</tr>
<tr>
<td>6</td>
<td>Artl_Dist (KM)</td>
<td>3.89</td>
<td>4.98</td>
<td>Distance to the nearest arterial line</td>
</tr>
<tr>
<td>7</td>
<td>AM_Service</td>
<td>7</td>
<td>29</td>
<td>AM transit availability, average per grid cell of 750 meters by 750 meters</td>
</tr>
<tr>
<td>8</td>
<td>Neighbor_Jobs</td>
<td>787</td>
<td>2715</td>
<td>Number of job per surrounding grid cell of 750 m by 750 m in the neighbor TAZs</td>
</tr>
<tr>
<td>9</td>
<td>LandValue ($)</td>
<td>84006</td>
<td>66527</td>
<td>Adjusted average property value in the last TAZ</td>
</tr>
<tr>
<td>10</td>
<td>White</td>
<td>0.88</td>
<td>0.10</td>
<td>Portion of white ethnicity to the total population of last TAZ</td>
</tr>
<tr>
<td>11</td>
<td>Black</td>
<td>0.04</td>
<td>0.07</td>
<td>Portion of African American ethnicity to the total population of last TAZ</td>
</tr>
<tr>
<td>12</td>
<td>MedianIncome ($)</td>
<td>34324</td>
<td>10018</td>
<td>Median of income in the last year TAZ</td>
</tr>
<tr>
<td>13</td>
<td>Distance_Old</td>
<td>9.64</td>
<td>7.91</td>
<td>Last year commute distance</td>
</tr>
<tr>
<td>14</td>
<td>Distance_New</td>
<td>11.26</td>
<td>8.90</td>
<td>Current year commute distance</td>
</tr>
</tbody>
</table>

The results for parameter estimation of the choice set formation model (in which average work distance is considered the determining criteria for alternative filtering) are presented in Table 2. Model parameters are estimated by maximizing the likelihood function presented in equations (8)–(11) using the \textit{nlp} procedure provided by SAS 9.1.3 package. Before evaluating the quality of the estimated parameters, it should be noted that the effect of covariates in a hazard model is facilitated by incorporating a negative sign for parameters in formulation. In other words, if the
Figure 1. Distribution of the households in the Seattle Metropolitan Area; Left: New pattern of distribution after moves; Right: Old pattern of distribution before moves.

Table 2. Modeling Results for the Hazard Model

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Weibull Estimate</th>
<th>t value</th>
<th>Log-logistic Estimate</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamma/Alpha</td>
<td>1.47</td>
<td>31.81</td>
<td>97.97</td>
<td>1.09</td>
</tr>
<tr>
<td>Beta</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Constant</td>
<td>5.08</td>
<td>5.79</td>
<td>—1.83</td>
<td>—1.25</td>
</tr>
<tr>
<td>OldDistance</td>
<td>0.06</td>
<td>8.41</td>
<td>0.06</td>
<td>8.54</td>
</tr>
<tr>
<td>GasPrice</td>
<td>−2.23</td>
<td>−2.59</td>
<td>−2.20</td>
<td>−2.57</td>
</tr>
<tr>
<td>Mortgage</td>
<td>0.28</td>
<td>2.28</td>
<td>0.28</td>
<td>2.29</td>
</tr>
<tr>
<td>Tot 6.17</td>
<td>−0.09</td>
<td>−1.75</td>
<td>−0.09</td>
<td>−1.73</td>
</tr>
<tr>
<td>Tot Adult</td>
<td>0.26</td>
<td>3.12</td>
<td>0.26</td>
<td>3.13</td>
</tr>
<tr>
<td>Employed</td>
<td>−0.13</td>
<td>−2.06</td>
<td>−0.13</td>
<td>−2.08</td>
</tr>
<tr>
<td>Ard. Dist (KM)</td>
<td>−1.82</td>
<td>−2.66</td>
<td>−1.83</td>
<td>−2.66</td>
</tr>
<tr>
<td>AM_Service</td>
<td>−0.0035</td>
<td>−1.83</td>
<td>−0.0035</td>
<td>−1.83</td>
</tr>
<tr>
<td>Neighbor Jobs (×100,000)</td>
<td>−2.92</td>
<td>−1.24</td>
<td>−2.96</td>
<td>−1.25</td>
</tr>
<tr>
<td>White</td>
<td>−0.68</td>
<td>−2.06</td>
<td>−0.70</td>
<td>−2.14</td>
</tr>
<tr>
<td>MedianIncome (×$100,000)</td>
<td>1.18</td>
<td>2.77</td>
<td>1.20</td>
<td>2.82</td>
</tr>
<tr>
<td>LandValue (×$100,000)</td>
<td>−0.12</td>
<td>−3.80</td>
<td>−0.11</td>
<td>−3.76</td>
</tr>
<tr>
<td>Likelihood value with only constant</td>
<td>−2318.45</td>
<td>446.80</td>
<td>445.32</td>
<td></td>
</tr>
<tr>
<td>Likelihood value at convergence</td>
<td>−2095.05</td>
<td>446.80</td>
<td>445.32</td>
<td></td>
</tr>
</tbody>
</table>
parameter of a covariate receives a negative sign, the chance of failure or the probability of accepting a work distance is increased. Alternatively, having a positive sign means that any increase in the covariate value decreases the chance of failure for the household, which implies that the household tends to increase the work distance.

Table 2 shows the parameter estimation results for the parametric hazard function with two different baseline hazard functions. Almost all of the covariates are relatively very close in both models. The Weibull distribution of the work distance model has a monotonically increasing shape because the gamma parameter is found to be greater than one. The beta parameter of the log-logistic distribution of the commute distance model is greater than one, meaning that it has a nonmonotonic function. Nonetheless, the alpha parameter is so big that it results in a monotonically increasing pattern, like the Weibull model, in a meaningful range of work distances (see Figure 2). Therefore, there is no advantage in using the log-logistic distribution because, with one extra variable, it neither improves the objective function nor provides a nonmonotonic pattern. It can be concluded from the baseline hazard analysis that, in the context of work distance, as the average commute distance increases, the hazard of moving to a closer residence increases. This is also completely in line with the intuitive feeling that people prefer to select residences that are closer to their job locations. Figure 2 shows the pattern of baseline hazard function for the two baseline function types.

As expected, the old_distance variable, average commute distance of last year, is included in the model with a high t-statistic, which to some extent means households keep their attitude of residing in the same distance from job location when they move to a new location. Gas price has a negative sign in the model, suggesting that, during
years with high gas price, movers give higher priority to housing alternatives that are closer on average to their job locations. This very policy-sensitive variable can affect the urban sprawl and reduce emission by reducing Vehicle Miles Traveled (VMT). Another macroeconomic-related variable included in the modeling analysis of this article is the mortgage rate. It was found that, as mortgage rate increases, households also increase their commute distance, and they consider residential locations that are not necessarily close to their job location.

The total number of children of ages 6 to 17 years old in the household is another critical variable in the models of this study. Households with more children of this age cohort select their houses relatively in closer locations to their jobs. This proximity gives a household the chance to provide a ride to their children who probably attend nearby schools. Unlike the total number of children, as the total number of adults increases, households may consider housing units with longer commute distances. This is only true when the adults are not all employed because the sign of the employed variable is negative, indicating that having more employed members in a household restricts their residential location search to only searching closer to their job locations.

Several built-environment variables have found statistically significant results in the modeling, shown in Table 3. The first variable among these land-use variables is

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gamma</td>
<td>1.87</td>
<td>14.24</td>
</tr>
<tr>
<td>Sigma</td>
<td>-0.62</td>
<td>-7.91</td>
</tr>
<tr>
<td>Constant</td>
<td>5.69</td>
<td>5.32</td>
</tr>
<tr>
<td>Old_Distance</td>
<td>0.09</td>
<td>6.04</td>
</tr>
<tr>
<td>GasPrice</td>
<td>-2.66</td>
<td>-2.58</td>
</tr>
<tr>
<td>Mortgage</td>
<td>0.38</td>
<td>2.53</td>
</tr>
<tr>
<td>Tot_6_17</td>
<td>-0.07</td>
<td>-1.15</td>
</tr>
<tr>
<td>Tot_Alt</td>
<td>0.27</td>
<td>2.84</td>
</tr>
<tr>
<td>Employed</td>
<td>-0.11</td>
<td>-1.44</td>
</tr>
<tr>
<td>Aver_Dist (KM)</td>
<td>-1.96</td>
<td>-2.26</td>
</tr>
<tr>
<td>AM_Service</td>
<td>-0.0034</td>
<td>-1.04</td>
</tr>
<tr>
<td>Neighbor_Jobs (×100,000)</td>
<td>-3.97</td>
<td>-1.10</td>
</tr>
<tr>
<td>White</td>
<td>-1.25</td>
<td>-2.54</td>
</tr>
<tr>
<td>Median Income (×$100,000)</td>
<td>1.65</td>
<td>2.96</td>
</tr>
<tr>
<td>LandValue (×$100,000)</td>
<td>-0.12</td>
<td>-2.96</td>
</tr>
<tr>
<td>Likelihood value with only constant</td>
<td>-2318.45</td>
<td></td>
</tr>
<tr>
<td>Likelihood value at convergence</td>
<td>-2082.99</td>
<td></td>
</tr>
<tr>
<td>( -2 \left[ L(C) - L(b) \right] )</td>
<td>470.92</td>
<td></td>
</tr>
</tbody>
</table>
the distance to the nearest arterial line with a negative sign. It can be interpreted from this variable that households living in dense urban areas with higher accessibility to the transportation network prefer to reside closer to their job locations. These types of people have a higher level of dependency on public transportation services (AM_Service) and therefore select residential locations that are covered by transit network. The next utilized built-environment variable included in the model was the total number of jobs in the surrounding zones. Households living in zones with more jobs available in the surrounding districts likely remain in zones where job opportunity is high and may remain high for years.

There are two land-use and demographic variables that were included in the model of this study; namely White and MedianIncome (median of income of the TAZ). In the Seattle region, households—living in zones in which the portion of people with Caucasian ethnicity is higher—tend to dwell closer to their job locations, whereas households in areas with higher-income levels prefer to move to places which are not necessarily close to their job locations.

The last variable included in the model was the property value of the current residence in which the household resides. As the average property price in the current zone (where the household resides) increases, it becomes more likely that the household moves closer to the employed members’ job locations. From another perspective, households with cheaper residences consider a larger pool of alternatives when they search for a new house, and they may consider residential locations that are farther from their job locations.

The estimated coefficients discussed in the above paragraphs can be used in equations (8)–(11) to simulate the expected work distance for each household. This simulation analysis is discussed in the next section.

**Simulation and Validation**

The cumulative density function discussed in equation (2) is used here to randomly simulate a work distance for each household using the results discussed in the previous section. Figure 3 shows a schematic comparison between the simulated results for the estimated models for two hazard models with alternative baseline hazard functions and the observed values.

Figure 3 shows that both hazard functions with Weibull and log-logistic baseline functions provide relatively similar patterns. The general pattern of the three curves is comparable as well.

Although the general pattern of the simulated results and the observed result are comparable, a gap can be seen between the simulated and observed results when work distance is less than 15 km. It is proposed in this article that this gap can be partially reduced if an observed heterogeneity variable is included in the hazard function. This random variable can capture the impact of the unobserved factors that have not been captured by the covariates used in the previous model. Further
discussion about including a random, unobserved heterogeneity variable in the hazard model is presented in the next section.

**Hazard-based Model with Unobserved Heterogeneity**

Unobserved heterogeneity is included in a duration model when there are unobserved effects, other than covariates that influence durations. Unobserved heterogeneity is typically controlled using a parametric random distribution (Flinn and Heckman 1982). A gamma distribution is the most widely used parametric distribution for correcting unmeasured heterogeneity (Lancaster 1979).

As discussed in the previous section, the nonmonotonic pattern of the Weibull distribution provides a better fit to the data of this study. Therefore, in this section of article, the hazard formulation with unobserved heterogeneity is only developed.

![Figure 3. Validation results for comparison between the observed and simulated results.](image)
with the Weibull baseline hazard. If a gamma distribution with a mean of one and variance of \(1/\sigma^2\) is considered for the exponential distribution of the unobserved heterogeneity variable, then the survival and PDF with the Weibull baseline hazard function can be formulated as:

\[
S_i(t) = \left[1 + \sigma^2 \times t^\gamma \times e^{\beta_i} \right]^{1/\delta}
\]

\[
f_i(t) = \gamma t^{\gamma-1} \times e^{\beta_i} \left[1 + \sigma^2 \times t^\gamma \times e^{\beta_i} \right]^{1/\delta-1},
\]

where the definition of the parameters are as those presented in the fourth section. The above equations are estimated using the Moment-Generating Function of a gamma random variable. Using the above equations, the model parameters are estimated the same way they were estimated in the last section. Results of this estimation are reported in Table 3.

As seen in Table 3, all the variables are relatively close to their corresponding parameters presented in Table 2. Therefore, we refrain from discussing them again in this section. The sigma parameter is statistically significant, which means that the unobserved heterogeneity exists in the model, and a gamma distribution can represent it in a statistically reasonable way.

In order to compare the goodness of fit of Table 3 with the models of Table 2, the Bayesian Information Criteria (BIC) is employed. BIC is a criterion for model selection among a set of parametric models with different numbers of parameters. It should be noted that a lower BIC implies a better model fit. The BIC formulation is

\[
\text{BIC} = -\ln(L_c) + 0.5p \ln(N),
\]

where \(\ln(L_c)\) is the log-likelihood value, \(p\) is the number of parameters, and \(N\) is the sample size.

The BIC value for the model with the Weibull baseline hazard function and without heterogeneity is calculated to be 2140.787, which is equal to 2133.99 for the model with heterogeneity. Therefore, the model with heterogeneity and a smaller BIC provides better fit for the data. Furthermore, if we compare the two models using the \(-2[L(C)_{\text{with Heterogeneity}} - L(C)_{\text{without Heterogeneity}}]\) statistic with the degree of freedom of one, it is found that the value of this statistic is 24.12. This amount of improvement by the model with heterogeneity is statistically meaningful at the .001 significance level.

In short, it can be said that including heterogeneity in the parametric hazard formulation with the Weibull baseline hazard function provides a better model for the work distance data.

Finally, a validation analysis similar to what was presented in the previous section is conducted in this section, also using the estimated parameters shown in Table 3. Figure 4 shows the comparison between the distribution of the observed work distances in the Seattle area and the distribution of the simulated data using the model with heterogeneity.
As seen in Figure 4, the general pattern of the new model with heterogeneity seems to be closer to the observed data than the models without heterogeneity in Figure 3.

**Conclusions and Future Directions**

This study presented a behavioral model of alternative set formation for residential location choice problem. In a housing search process, one can consider a two-step approach in which alternatives are evaluated and screened based on household priorities, lifestyle, and preferences and for each alternative. The probability of being selected in the choice set is estimated. Following that, the alternative with the highest utility can be selected using traditional choice models. This study focused on the first step of the household housing search process. However, the choice set formation methodology that is introduced in this article can then be coupled with a discrete choice model to present a comprehensive location choice model.

The housing screening process of this study is modeled using average work distance as a continuous variable. The analogy between distance and duration implied the application of hazard-based formulation to model the willingness of the employed household members to accept a commute distance. By contributing to the
literature of the recently introduced spatial hazard models, this study explored and discussed a new application of spatial hazard-based models for housing search behavior modeling. We also discussed some of the specifications of the hazard-based models that are less frequently utilized, even in duration applications such as the nonmonotonic baseline hazard function of log-logistic functions and the parametric unobserved heterogeneity of gamma distribution. Introduction of these specifications is a positive addition to the literature of spatial hazard models.

The PSTP of Seattle Metropolitan Area was used in this study for the modeling practice, along with other sources of data, such as built-environment, land-use, and economic factors. Many household sociodemographic attributes and several land-use indicators were tested in the modeling process, among which the total number of youths, adults, employed members, the property value, and the income were found to be statistically significant in the model. Among built-environment variables, the morning transit availability and total job counts in the region were significant in the final model. From the supply side of the market, inflation rate, mortgage rate, and gas prices were proved to influence the household residential location decision. It was shown that, by including a random, unobserved heterogeneity variable among the covariates, the general goodness of fit of the model can be significantly improved.

Further improvements to the model include investigating the importance of variables (beyond work distance) on housing search choice set formation and integrating the choice set probabilities with a discrete choice model. These improvements remain as future research tasks. It should also be noted that the application of the proposed modeling framework is not limited to the housing search problem. Such a framework can be used in other contexts in which a large number of alternatives should be evaluated. For instance, in the case of activity location choice (e.g. shopping), a similar approach can be used.

Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

References
Movement as Complex Human Behavior.” Transportation Research Record 723:11–24.
———. 1982. “The Analysis of Travel as an Example of Complex Human Behavior in
Spatially-Constrained Situations: Definitions and Measurement Issues.” Transportation

Casetta, E., F. Pagliara, and K. Axhausen. 2007 “The Use of Dominance Variables in Choice
Set Generation.” In Proceedings of the 11th World Conference on Transport Research,
University of California at Berkeley.

Hazard Analysis of Urban Form in American Metropolitan Areas.” Papers in Regional

Cox, D. R. 1959. “The Analysis of Exponentially Distributed Life-time with Two Types of


Amsterdam, The Netherlands: North-Holland.


Studies in Regional Science and Urban Economics. North Holland, The Netherlands:
Elsevier.


Han, A., and J. A. Hausman. 1990. “Flexible Parametric Estimation of Duration and


Image and Constraints on Destination Choice: An Application of the NMNL Model.”

King County Assessment Department. 2009. “Parcel Level Property Values, King

Kim, S. 1992. “Search, Hedonic Prices and Housing Demand.” The Review of Economics and


